Project 1 - Thương mại điện tử (E-commerce A/B Testing)

Nhóm 15

2024-10-28

Họ và tên thành viên

1. Đậu Quang Anh - 22110014
2. Lâm Gia Bảo - 22110023
3. Trần Quốc Danh - 22110035
4. Lê Thị Hồng Đang - 22110033
5. Trần Duy An - 22110008

Bài làm

```
data <- read_csv(file = "D:/XLSLTK/datasets/ab_test_commerce.csv",na = c("", "NA", "N/A"))
```

```
## Rows: 294478 Columns: 6
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (3): date, group, landing_page
## dbl  (2): user_id, converted
## time (1): timestamp
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
data <- data |> clean_names()
glimpse(data)
```

```
## Rows: 294,478
## Columns: 6
## $ user_id      <dbl> 851104, 804228, 661590, 853541, 864975, 936923, 679687, 7~
## $ date         <chr> "2017-1-21", "2017-1-12", "2017-1-11", "2017-1-8", "2017-~
## $ timestamp    <time> 22:11:49, 08:01:45, 16:55:06, 18:28:03, 01:52:26, 15:20:~
## $ group        <chr> "control", "control", "treatment", "treatment", "control"~
## $ landing_page <chr> "old_page", "old_page", "new_page", "new_page", "old_page~
## $ converted    <dbl> 0, 0, 0, 0, 1, 0, 1, 0, 1, 1, 0, 0, 0, 0, 0, 1, 0, 1, 0, ~
```

```
data |> group_by(user_id)
```

```
## # A tibble: 294,478 x 6
## # Groups:   user_id [290,584]
##    user_id date      timestamp group     landing_page converted
##      <dbl> <chr>     <time>    <chr>     <chr>            <dbl>
##  1  851104 2017-1-21 22:11:49  control   old_page             0
##  2  804228 2017-1-12 08:01:45  control   old_page             0
##  3  661590 2017-1-11 16:55:06  treatment new_page             0
##  4  853541 2017-1-8  18:28:03  treatment new_page             0
##  5  864975 2017-1-21 01:52:26  control   old_page             1
##  6  936923 2017-1-10 15:20:49  control   old_page             0
##  7  679687 2017-1-19 03:26:47  treatment new_page             1
##  8  719014 2017-1-17 01:48:30  control   old_page             0
##  9  817355 2017-1-4  17:58:09  treatment new_page             1
## 10  839785 2017-1-15 18:11:07  treatment new_page             1
## # i 294,468 more rows
```

```
data_country <- read_csv(file = "D:/XLSLTK/datasets/countries_ab_test_commerce.csv", na = c("", "NA", "
```

```
## Rows: 290584 Columns: 2
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): country
## dbl (1): user_id
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
data_country <- data_country |> clean_names()
glimpse(data_country)
```

```
## Rows: 290,584
## Columns: 2
## $ user_id <dbl> 834778, 928468, 822059, 711597, 710616, 909908, 811617, 938122~
## $ country <chr> "UK", "US", "UK", "UK", "UK", "UK", "US", "US", "US", "US", "U~
```

1. Trong dữ liệu này có một số lượng nhất định người dùng đã thực hiện nhiều hơn 1 lần tương tác với trang web của công ty (cả cũ và mới), do đó, cần hiệu chỉnh/làm sạch dữ liệu trước khi phân tích. Hãy viết ra lựa chọn xử lý và thực hiện trên đoạn code chương trình.

Gộp dữ liệu của data_ab_test và data_countries theo cột user_id thành data_merged

```
data_ab <- merge(data,data_country,by="user_id",all.x=TRUE)

glimpse(data_ab)
```

```
## Rows: 294,478
## Columns: 7
## $ user_id      <dbl> 630000, 630001, 630002, 630003, 630004, 630005, 630006, 6~
## $ date         <chr> "2017-1-19", "2017-1-16", "2017-1-19", "2017-1-12", "2017~
## $ timestamp    <time> 06:26:07, 03:16:43, 19:20:56, 10:09:32, 20:23:59, 21:22:~
## $ group        <chr> "treatment", "treatment", "control", "treatment", "treatm~
## $ landing_page <chr> "new_page", "new_page", "old_page", "new_page", "new_page~
## $ converted    <dbl> 0, 1, 0, 0, 0, 1, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, ~
## $ country      <chr> "US", "US", "US", "US", "US", "US", "US", "US", "US", "UK~
```

Chuyển đổi timestamp

```
#data_ab <- data_ab |> mutate(time_hour = hour(timestamp),
 #                            time_minutes = minute(timestamp),
  #                           time_second = second(timestamp))
# glimpse(data_ab)


data_ab <- data_ab |> mutate(time=hour(timestamp)*60+minute(timestamp)+second(timestamp))
glimpse(data_ab)
```

```
## Rows: 294,478
## Columns: 8
## $ user_id      <dbl> 630000, 630001, 630002, 630003, 630004, 630005, 630006, 6~
## $ date         <chr> "2017-1-19", "2017-1-16", "2017-1-19", "2017-1-12", "2017~
## $ timestamp    <time> 06:26:07, 03:16:43, 19:20:56, 10:09:32, 20:23:59, 21:22:~
## $ group        <chr> "treatment", "treatment", "control", "treatment", "treatm~
## $ landing_page <chr> "new_page", "new_page", "old_page", "new_page", "new_page~
## $ converted    <dbl> 0, 1, 0, 0, 0, 1, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, ~
## $ country      <chr> "US", "US", "US", "US", "US", "US", "US", "US", "US", "UK~
## $ time         <dbl> 393, 239, 1216, 641, 1282, 1308, 365, 982, 1431, 1002, 58~
```

Hiệu chỉnh dữ liệu bị sai khi group = control và landing_page = new_page, group = treatment và landing_page = old_page

```
data_ab|>filter(group == "control" & landing_page != "old_page") |> nrow()
```

```
## [1] 1928
```

```
data_ab|>filter(group == "treatment" & landing_page != "new_page") |> nrow()
```

```
## [1] 1965
```

```
data_clean <- data_ab |>  mutate(landing_page = case_when(
  group == "control" & landing_page != "old_page" ~ "old_page",
  group == "treatment" & landing_page != "new_page" ~ "new_page",
  TRUE ~ landing_page ))
data_clean|>filter(group == "control" & landing_page != "old_page") |> nrow()
```

```
## [1] 0
```

```
data_clean|>filter(group == "treatment" & landing_page != "new_page") |> nrow()
```

```
## [1] 0
```

```
glimpse(data_clean)
```

```
## Rows: 294,478
## Columns: 8
```

```
## $ user_id      <dbl> 630000, 630001, 630002, 630003, 630004, 630005, 630006, 6~
## $ date         <chr> "2017-1-19", "2017-1-16", "2017-1-19", "2017-1-12", "2017~
## $ timestamp    <time> 06:26:07, 03:16:43, 19:20:56, 10:09:32, 20:23:59, 21:22:~
## $ group        <chr> "treatment", "treatment", "control", "treatment", "treatm~
## $ landing_page <chr> "new_page", "new_page", "old_page", "new_page", "new_page~
## $ converted    <dbl> 0, 1, 0, 0, 0, 1, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, ~
## $ country      <chr> "US", "US", "US", "US", "US", "US", "US", "US", "US", "UK~
## $ time         <dbl> 393, 239, 1216, 641, 1282, 1308, 365, 982, 1431, 1002, 58~
```

```
sum(duplicated(data_clean$user_id))
```

```
## [1] 3894
```

Như vậy thì có 3894 user_id bị trùng lặp nên ta sẽ loại bỏ những user_id này bằng hàm `distinct()`

```
data_clean <- data_clean |> distinct(user_id, .keep_all = TRUE)
```

2. Bảng tóm tắt, khái quát về dữ liệu.

```
converted_summarise <- data_clean |> group_by(landing_page, converted) |>
  summarise(n = n(), tb =mean(time), sd = sd(time))
```

```
## `summarise()` has grouped output by 'landing_page'. You can override using the
## `.groups` argument.
```

```
converted_summarise
```

```
## # A tibble: 4 x 5
## # Groups:   landing_page [2]
##   landing_page converted      n    tb    sd
##   <chr>            <dbl>  <int> <dbl> <dbl>
## 1 new_page             0 128085  748.  416.
## 2 new_page             1  17271  758.  415.
## 3 old_page             0 127764  750.  416.
## 4 old_page             1  17464  751.  415.
```

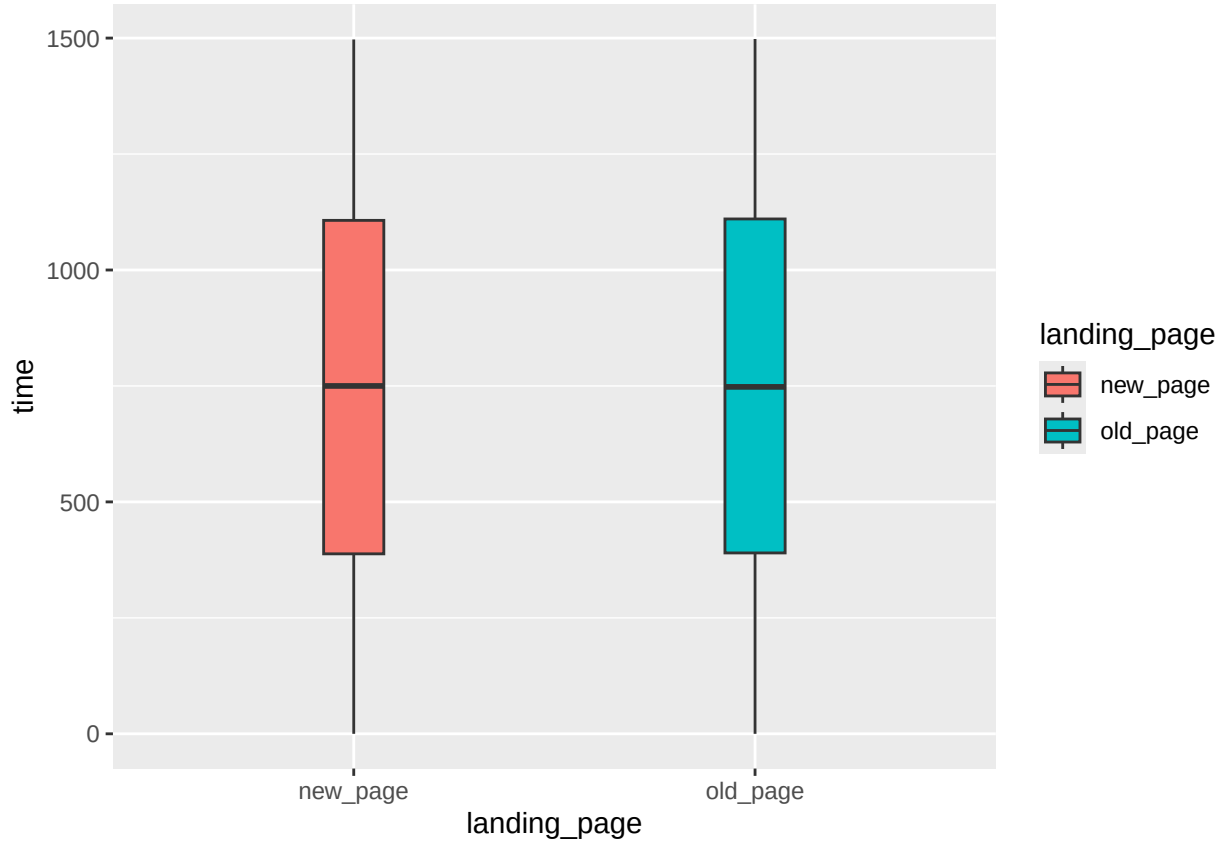
```
summarises <- data_clean |> group_by(landing_page) |>
  summarise(n = n(), tb = mean(time), sd = sd(time))
summarises
```

```
## # A tibble: 2 x 4
##   landing_page      n    tb    sd
##   <chr>         <int> <dbl> <dbl>
## 1 new_page     145356  749.  416.
## 2 old_page     145228  750.  416.
```

Kết quả cho thấy số lượng người dùng truy cập vào trang web cũ và trang web mới là tương đương nhau. Tuy nhiên, thời gian truy cập trung bình của trang web mới cao hơn trang web cũ.

Biểu đồ boxplot cho thời gian truy cập trung bình của 2 trang web.

```
ggplot(data_clean, aes(x=landing_page, y=time, fill=landing_page)) +
  geom_boxplot(width = 0.15)
```



3. Đề ra các phương án xử lý dữ liệu dựa trên các công cụ của A/B testing nhằm đưa ra bằng chứng để trả lời cho câu hỏi: "Trang web mới có thực sự tốt hơn trang web cũ?".

Vì ta cần kiểm tra trang web mới có tốt hơn trang web cũ hay không nên đối thuyết H1 sẽ là p1 < p2

Tỷ lệ người mua ở trang web mới

```
new_page_rate <- converted_summarise$n[converted_summarise$landing_page=="new_page" &
                                       converted_summarise$converted==1]/summarises$n[summarises$landi
new_page_rate
```

```
## [1] 0.1188186
```

Tỷ lệ người mua ở trang web cũ

```
old_page_rate <- converted_summarise$n[converted_summarise$landing_page=="old_page" &
old_page_rate
```

```
## [1] 0.1202523
```

Gọi p_1, p_2 lần lượt là tỷ lệ mẫu người mua ở trang web cũ (old_page_rate) và trang web mới (new_page_rate)

Ta có được với mức ý nghĩa là $\alpha = 0.05$.

Giả thuyết $H_0 : p_1 = p_2$

Đối thuyết $H_1 : p_1 < p_2$

Nếu Giả thuyết là đúng thì sự nhiều hơn về tỷ lệ người mua ở trang web mới so với trang web cũ chỉ là kết quả của sự ngẫu nhiên (không có ý nghĩa thống kê). Ngược lại, nếu Đối thuyết là đúng thì sự nhiều hơn về tỷ lệ người mua ở trang web mới so với trang web cũ là có ý nghĩa thống kê.

Ta sẽ sử dụng phương pháp Permutation Test

4. Hãy cố gắng tận dụng hết các biến được cung cấp, để xử lý dữ liệu theo các phương án đã đề ra.

```
set.seed(21)
rate_perm_fun <- function(x, y, R, p_A, p_B, alter){
  data <- split(x,y)
  n <- length(x)
  nA <- length(data[[1]])
  nB <- length(data[[2]])
  mean_diff <- numeric(R)
  for (i in 1:R){
    idx_a <- sample(x = 1:n, size = nA)
    idx_b <- setdiff(x = 1:n, y = idx_a)
    mean_diff[i] <- sum(x[idx_a])/nA-sum(x[idx_b])/nB
  }
  if (alter == "left_sided"){
    p_values <- mean(mean_diff < (p_A-p_B))
  }
  else if (alter == "right_sided"){
    p_values <- mean(mean_diff > (p_A-p_B))
  }
  else{
    p_values <- mean(abs(mean_diff) > (p_A-p_B))
  }
  return (list(res_perm = mean_diff,rate_diff = p_A-p_B, p_value = p_values))
}

result <- rate_perm_fun(data_clean$converted,data_clean$landing_page, R = 1000,
                        p_A=old_page_rate, p_B = new_page_rate,alter = "left_sided")

result$p_value
```

```
## [1] 0.886
```

Vì p-value $= 0.886 >$ alpha $= 0.05$ nên chưa đủ cơ sở bác bỏ H_0. Vậy tỷ lệ người mua giữa hai trang web không có ý nghĩa thống kê, tức trang web mới không hiệu quả nhiều so với trang web cũ.

5. Viết các nhận xét và kết luận

Phân tích cho thấy thời gian truy cập giữa hai trang web là tương đương. Tuy nhiên, tỷ lệ người dùng trả tiền trên trang web mới không cao hơn trang web cũ.

Kết luận: Công ty không cần triển khai trang web mới, nên giữ lại trang cũ và tập trung vào các chiến lược khác để tăng doanh thu.